Task-Oriented Uncertainty Evaluation for Linked Data Based on Graph Interlinks

Ahmed El Amine Djebri, Andrea G. B. Tettamanzi, Fabien Gandon

Task-Oriented Uncertainty Evaluation for Linked Data Based on Graph Interlinks

Ahmed El Amine Djebri[0000−0003−2917−5085], Andrea G.B. Tettamanzi[0000−0002−8877−4654], and Fabien Gandon[0000−0003−0543−1232]

Université Côte d'Azur, Inria, CNRS, I3S, France

Abstract. For data sources to ensure providing reliable linked data, they need to indicate information about the (un)certainty of their data based on the views of their consumers. In Addition, uncertainty information in terms of Semantic Web has also to be encoded into a readable, publishable, and exchangeable format to increase the interoperability of systems. This paper introduces a novel approach to evaluate the uncertainty of data in an RDF dataset based on its links with other datasets. We propose to evaluate uncertainty for sets of statements related to user-selected resources by exploiting their similarity interlinks with external resources. Our data-driven approach translates each interlink into a set of links referring to the position of a target dataset from a reference dataset, based on both object and predicate similarities. We show how our approach can be implemented and present an evaluation with real-world datasets. Finally, we discuss updating the publishable uncertainty values.

Keywords: Uncertainty · Semantic Web · Graph interlinks.

1 Introduction

We are witnessing an era fulfilling the vision to create a Web of linked intelligent systems [1], thriving through sharing data they own or have processed. In this context, many challenges present themselves to developers of such platforms to retain reliable data that allows enriching their existing knowledge bases using robust reasoning or with the help of more external relevant content. The latter is using links with extra pieces of information revealing new dimensions for users to explore with their requests. Uncertainty is a major issue when related to content brought out on the Web, or Semantic Web by extension. Nevertheless, most data providers do not present explicit information about the uncertainty of their data. On the other hand, completely mistrusting a data source is unfair: while some data providers may not be reliable on one subject or provide false information about it, they are experts on other subjects and the pieces of information they provide should not be ignored. In some cases, references about data provenance and/or related data are given, from which a data consumer may hope to get further validation from other data sources.

In this paper, we address the need to evaluate uncertainty in linked data sources. In our approach, a data source may auto-evaluate the level of uncertainty of its data according to what is being presented by other data sources and for a specific use-case. We leverage the fact that different knowledge graphs may provide complementary and/or extra information enabling the assessment of the conformity of a target source. We also think that a user's preferences should be taken into consideration while evaluating uncertainty. Our work is built on top of the *mUnc* model [2] to represent and publish uncertainty on the Semantic Web. The main question we aim to answer is: *How to evaluate uncertainty in a data source, based on its data, other linked data sources, and with respect to a specific use-case?*

To answer this question, we propose an approach to evaluate the uncertainty of a target data source, based on graph interlinks with other reference data sources. We propose to annotate statements with uncertainty values in a publishable format and provide a method to manipulate and update such values if existed. The intuition behind this work is that often users who need to confirm a piece of information will look for different sources that confirm or contradict it. For instance, the traditional verification techniques in journalism include the "two-sources rule" asking to verify that at least two independent trustworthy sources confirm a piece of information.

The rest of the paper is organized as follows. Section 2 surveys related work and positions our contribution accordingly. In section 3 we discuss similarity assessment between two focus graphs of one resource and our choices of indicators. In section 4 we present our main contribution, with a method to evaluate uncertainty based on existing links and transform it into reusable information that annotates statements in the data source of interest. In section 5 we discuss the experimental workflow and present our tool for uncertainty evaluation and annotation. We conclude with a snapshot of our work and our future goals.

2 Related Works

According to Paulheim [3], external error-detection approaches in knowledge graphs are based on interconnections between data sources: they take advantage of the links (identity links or simply IRI reuse) to check for errors in the data source of interest. Paulheim [4] proposes in another work an external approach to detect outlier interlinks between datasets by creating a feature vector representation of each interlink based on types and incoming/outgoing links to all instances of a class. That work is meant to evaluate links, whilst here we check the reliability of data based on presumed correct interlinks. Other works are based on a statistical analysis of feature vectors associated with predicates that are linked to interlinked resources [5,6]. Another interesting idea is identity quantification between two linked data sets. It explores the idea of isomorphism quantification between two sets presumably representative of the same real-world entity. Similar works inspiring data-driven ontology alignment were discussed by Shvaiko *et al.* [7].

Christodoulou *et al.* [8] discusses the use of similarity measurements and Bayesian updating to help to align ontologies from different data sources and using precomputed values provided by ontology matchers. The authors depend on the Linked Open Vocabularies[1] to calculate the likelihood of equivalence vs. non-equivalence of two distinct classes and use that measure to update the local probability of similarity between two classes using Bayesian update. Authors of [9] propose a statistical data-driven approach to detect incorrect property mappings among the different language chapters of DBpedia. The work focuses on detecting the wrong mappings and the analysis is run through the whole datasets.

The aforementioned works mostly treated the reliability of the similarity links between data sources or detecting wrong schema-mappings. Different from our problem that requires analyzing data based on a use-case. The previous works present a promising set of measures to analyze data uncertainty based on links. Nevertheless, we notice the absence of specific sets of interest encapsulating the linked resources. Moreover, the said works are more in the spirit of ontology-matching techniques relying on linking all instances of two classes.

The problem relates in general to ontology alignment approaches and is also inspired by quasi-key detection problems. Most of the literature is assessing the link quality and not depending on the links themselves to assess data quality. We believe that it is original to discuss uncertainty evaluation with a task-centered perspective based on graph interlinks .

3 Uncertainty Assessment in Linked Data

3.1 Terminology and Definitions for Uncertainty

We introduce the terminology and the formalism used in the paper to propose an evaluation of uncertainty based on existing links between graphs.

Definition 1. *RDF-dataset — a set of statements (triples) in the form* $\langle subject, predicate, object\rangle \in (I \cup B) \times I \times (I \cup B \cup L)$ *where* I *is a set of IRIs,* B *a set of blank nodes,* L *a set of literals,* I, B *and* L *are pairwise disjoint and for every two RDF-datasets* D_1, D_2 *the sets of blank nodes are disjoint. we also denote* I_D *the set of IRIs used in statements of the RDF-dataset* D*.*

Definition 2. *Target dataset — an RDF-dataset noted as* D_t *that is the target of the uncertainty evaluation.*

Definition 3. *Reference dataset — an RDF-dataset noted as* D_r *that represents a reference for the evaluation of the uncertainty of a target dataset.*

Definition 4. *RDF-graph — a graph* $G = (V, E)$*, where* $V \subset (I \cup B \cup L)$ *is a set of vertices, and* $E \subset I$ *is a set of directed edges.*

[1] https://lov.linkeddata.es/

Definition 5. *Focus graph — an RDF-graph noted as $G_D(e) \subset D$, where D is the dataset including the graph (target or reference) and $e \in I$ is a focused resource for which $G_D(e)$ is considered representative according to the use-case.*

Definition 6. *Set of Linking predicates — a non-empty set of predicates explicitly chosen to link between the target dataset and the reference dataset. We note it as $P_l \subset I$. Example: $P_l = \{$`owl:sameAs`, `skos:exactMatch`$\}$.*

Definition 7. *Contextual Linkset — as defined in the VOID[2] vocabulary, a linkset is a set of RDF triples where all subjects are in one dataset and all objects are in another dataset. We call a contextual linkset the one containing links between focused resources of D_t and those of D_r. A contextual linkset defines the set of focused resources of each dataset as well as the links between them. A link between a target focused resource e_t and a reference focused resource e_r is also a link between the focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$: $LS(D_t, D_r) = \{\langle e_t, p, e_r\rangle \mid p \in P_l, e_t \in I_{D_t}, e_r \in I_{D_r}\}$;*

Definition 8. *Evidence link — a relationship between two statements $t_t \in G_{D_t}(e_t)$, $t_r \in G_{D_t}(e_r)$ discovered using similarity analysis, that supports the link between two linked focus graphs $G_{D_t}(e_t)$ and $G_{D_t}(e_r)$. The evidence link refers to a considered relationship between the predicates and/or the objects of two statements t_t and t_r. We note $E(G_{D_t}(e_t), G_{D_t}(e_r))$ the set of evidence links discovered between the two focus graphs $G_{D_t}(e_t)$ and $G_{D_t}(e_r))$.*

Our purpose is to find a method to assess the reliability of the information in each target focus graph $G_{D_t}(e_t)$ centered around a target focused resource e_t. To this end, we translated the existing link between the resource e_t of a target dataset and the resource e_r of a reference dataset ($\langle e_t, p, e_r\rangle \in LS(G_{D_t}(e_t), G_{D_t}(e_r)), p \in P_l)$ to a set of evidence links between the target focus graph $G_{D_t}(e_t)$ and the reference focus graph $G_{D_t}(e_r)$. We statistically analyze the extracted evidence links to obtain a set of indicators enabling the evaluation of the overall semantic similarity between the predicates of linked focus graphs. Finally, we use the extracted evidence links to calculate the uncertainty of each focus graph based on its local ones.

3.2 Choosing Target Focused Resources

The problem of matching, whether it is data-driven or schema-driven, is context-related and may not be evident to users or useful for their request if done without involving them in the process [10]. We consider the concept of uncertainty to be also context-specific and that it is possible to choose a different evaluation method for each use case.

A focus graph $G_D(e)$ is meant to be the image that represents e in the context of the application. Hence, the choice of the set of focused resources is necessary to ensure that uncertainty assessment is built on a user-centered view. The set of

[2] https://www.w3.org/TR/void/#linkset

targeted focused resources $e \in I_{D_t}$ (I_{D_t} being the set of IRIs in the dataset D_t) depends on the type of validation a user intends to have within the data-source and depending on the use-case.

We also need to present a sufficient focus graph —in the context of the use-case—reflective of information about the resource. As an example with music artists, a focus graph may contain simple information like their names and birthplaces and deeper-level information like songs from their albums. To limit the issue in our current use-case, we rely on the proposal[3] made by Strikler, aiming to create a focused subgraph centered around and describing a resource, called a *Concise-Bounded Description* and noted as *CBD*. Some Linked Data stores like *Virtuoso*[4] propose their proper definition of *CBD* and use it as the mapping of `DESCRIBE` SPARQL queries. For our current use-case, we find the definition of *CBD* an intuitive, simple yet interesting one to define our $G_D(e)$.

3.3 Linking predicates and Contextual Linkset

Unlike the approaches to ontology matching or alignment, we take existing links in the contextual linkset as ground truth. The first links one may find between two data sources can be established by reusing IRIs of resources from one in the other. Moreover, the *RDFS* and *OWL* standards provide predicates such as `owl:sameAs`, `rdfs:seeAlso` with debatable semantics to link between resources [11,12]. Other commonly used ontologies propose more predicates to indicate the matching between two resources (example: `skos:exactMatch` [13]).

4 Approach and Uncertainty Assessment Pipeline

We propose a level-based architecture where each level depends on the previous one, from isolating candidate evidence links to exporting update-ready uncertainty values. A link between a target focused resource e_t and a reference focused resource e_r can be seen as a link between the focus graph of each. The evidence links supporting that link are discovered and selected based on defined similarity indicators.
Considering two statements $t_1 : \langle s_1, p_1, o_1 \rangle, t_2 : \langle s_2, p_2, o_2 \rangle$ where $t_1 \in G_{D_t}(e_t)$ and $t_2 \in G_{D_r}(e_r)$ and a prior knowledge indicating the existence of a link between the two resources e_t and e_r: $\langle e_t, l, e_r \rangle \in LS(D_t, D_r)$. We define here a set of similarity indicators to be used for uncertainty assessment.

4.1 Precomputing: Augmentation and Clustering

During this step, we apply the chosen definition of focus graphs on D_t based on $LS(D_t, D_r)$. Beforehand, we use *OWL* [12] semantics for properties to augment the data source by calculating the transitive closure of our target dataset D_t. This helps to unveil more potential evidence links between the linked focus graphs.

[3] https://www.w3.org/Submission/CBD/
[4] http://docs.openlinksw.com/virtuoso/rdfsqlfromsparqldescribe/

4.2 Level 1: Identifying possible evidence links based on Syntactic similarity between objects of statements in linked focus graphs

In the first level, we produce a set of evidence links for each pair of linked focus graphs using an *object similarity* measure defined as follows.

Definition 9. *Object similarity — We denote $sym_o(t_1, t_2)$ (eq. 1) as the weighted similarity between objects of statements t_1 and t_2 (between o_1 and o_2). This measure refers to what extent the two objects share the same nature (literal, URI), the same datatype(`xsd:short`, `xsd:integer`, etc.[5]) and/or the same value:*

$$sym_o(t_1, t_2) = (1 - \omega_{val}) \times typeMatch(o_1, o_2) + \omega_{val} \times valMatch(o_1, o_2). \quad (1)$$

The binary function $typeMatch$ returns 1 if both nature (IRI, Literal) and datatypes are similar and 0 otherwise. The $valMatch$ function can be any syntactic similarity measure (Jaccard, Levenshtein, Jaro-Winkler distance, n-grams, etc.). Once the first level measures are established, a positive threshold $\tau_{obj} \leq 1$ restricts the discovered evidence links to ones of higher object similarity.

4.3 Level 2: Identifying evidence link patterns using Semantic similarity of predicates in the overall linked focus graphs

The second level introduces semantic similarity between evidence links while taking into account: the fact that the same predicates are used in schemas of the different data sources, and specific semantics related to the current use case by the mean of predicate similarity indicators. This view is inspired by the example in [8] but adapted to fit predicates due to the generalized, class-independent definition of $LS(D_t, D_r)$.

Definition 10. *Predicate similarity — We denote $sym_p(t_1, t_2)$ (eq. 3) the statistical similarity between predicates of statements t_1 and t_2 (between p_1 and p_2). This measure is built on all the linked focus graphs and represents the use-case related semantic similarity of the two predicates p_1 and p_2.*

To evaluate semantic similarity, we first define four indicators (table 1) to be statistically extracted for each pair of linked focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$. From the previous indicators and for each pair of linked focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$, we also calculate three local ratios (table 2).

To evaluate the semantic similarity between p_1 and p_2 on the overall contextual linkset, we average for each pair of predicates p_1 and p_2 with an evidence link between t_1 and t_2 in all linked focus graphs, and add another indicator $\hat{R}_0$ for the equality $p_1 = p_2$ (as it will stay the same if averaged). We get a vector of averaged ratios $\hat{R}(p_1, p_2) = [\hat{R}_0(p_1, p_2), \hat{R}_1(p_1, p_2), \hat{R}_2(p_1, p_2), \hat{R}_3(p_1, p_2)]$, with

$$\hat{R}_i(p_1, p_2) = \frac{1}{|LS(D_t, D_r)|} \sum_{\langle e_t, p_l, e_r \rangle \in LS(D_t, D_r)} R_i(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2) \quad (2)$$

[5] https://www.w3.org/2011/rdf-wg/wiki/XSD_Datatypes

Table 1: Semantic similarity indicators for each pair of linked focus graphs.

Indicator	Definition
$I_1(G_{D_t}(e_t), G_{D_r}(e_r))$	the number of evidence links between the two focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$. i.e. the number of links supporting the similarity hypothesis between the two resources e_t and e_r.
$I_2(G_{D_t}(e_t), G_{D_r}(e_r))$	the set of predicate pairs in evidence links between statements of the two focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$. i.e: the set of pairs (p_1, p_2) where an evidence link exists between t_1 and t_2.
$I_3(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	the count of evidence links relying on two predicates p_1, p_2 between the two focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$.
$I_4(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	the total number of possible combinations between statements using each p_1 or p_2 in the two linked focus graphs $G_{D_t}(e_t), G_{D_r}(e_r)$ (For instance, if three statements in $G_{D_t}(e_t)$ use p_1 and two statements in $G_{D_r}(e_r)$ use p_2 then the total number of links would be six. So this represents the maximum possible number of evidence links that can be found linking p_1 and p_2).
$I_5(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	the sum of the quality of evidence links relying on two predicates p_1, p_2 between the two focus graphs $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$. i.e. the sum of object similarities of discovered evidence links between $G_{D_t}(e_t)$ and $G_{D_r}(e_r)$ linking statements using respectively p_1 and p_2.

Table 2: Normalised local ratios for each pair of linked focus graphs.

Ratio	Definition
$R_1(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	I_3 is normalised using I_1 to reflect the participation of evidence links between two statements having p_1 and p_2 as predicates, in the overall evidence links between the two linked focus graphs.
$R_2(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	I_3 is normalised using I_4 to reflect the portion of existing statement that actually participate with a link. If all existing statements between two focus graphs, with p_1 and p_2 as predicates are linked with evidence links, it indicates that the predicates may be functional, or that this information is a common knowledge that usually have a lower cardinal (like homepages for artists).
$R_3(G_{D_t}(e_t), G_{D_r}(e_r), p_1, p_2)$	I_5 is normalised using I_3 to get the average quality of each evidence link between statements having p_1 and p_2 as predicates.

and for which we define a vector of semantic weights $\omega_{sem} = [\omega_0, \omega_1, \omega_2, \omega_3]$ with $\sum \omega_i = 1, \omega_i \geq 0$. We select only the predicate pairs having an average of link quality equal or greater than a positive defined threshold τ_{sem} where $\tau_{sem} \leq \hat{R}_3(p_1, p_2) \leq 1$. Hence, we can define $sym_p(t_1, t_2)$ of statements t_1 and t_2 as the dot product of the two vectors $\hat{R}(p_1, p_2)$ and ω_{sem}:

$$sym_p(t_1, t_2) = \omega_{sem} \cdot \hat{R}(p_1, p_2) \tag{3}$$

Similarly to the previous level, the overall quality of considered evidence links should also respect the average quality threshold τ_{sem}.

4.4 Level 3: Evaluating Contextual Uncertainty of target focus graphs

At this level, the previous similarity measures are combined into one value reflecting the degree of uncertainty of a target focus graph $G_{D_t}(e_t)$ regarding its

linked reference focus graph $G_{D_r}(e_r)$. For this, we define the notion of contextual uncertainty to be the measure of one of a target focus graph based on its evidence links.

Definition 11. *Contextual Uncertainty — We define contextual uncertainty of a target focus graph $G_{D_t}(e_t)$ compared to a reference focus graph $G_{D_r}(e_r)$, with a link existing between e_t and e_r in the contextual linkset $LS(D_t, D_r)$, as the sum of products of object(syntactic) and predicate(semantic) similarity scores of the statements linked by each $l \in E(G_{D_t}(e_t), G_{D_r}(e_r))$, on the number of evidence links in $E(G_{D_t}(e_t), G_{D_r}(e_r))$.*

$$U(G_{D_t}(e_t) \mid \langle e_t, p_l, e_r \rangle) = \frac{\sum_{<t_1,l,t_2> \in E(G_{D_t}(e_t), G_{D_r}(e_r))} sym_o(t_1, t_2) \times sym_p(t_1, t_2)}{|E(G_{D_t}(e_t), G_{D_r}(e_r))|} \tag{4}$$

5 Experiment and Evaluation

We evaluate a dataset with 714 artists from *MusicBrainz* against their linked information from the English chapter of *dbpedia*. The used dataset including focus graphs and contextual linkset is available online.

To validate our approach, we developed *Archer*[6], a tool for analyzing and annotating link data with uncertainty values. *Archer* uses the proposed approach to extract the object and predicate similarity with respect to the links between focus graphs. The tool allows the user to query for identity links, extract focus graphs from both the target and the reference datasets, and evaluate the uncertainty of each focus graph in the target dataset. It further allows analyzing and visualizing pairs of linked focus graphs individually as well as the different indicators for the overall analysis.

For the experiment, we chose both a Jaccard distance and a string equality measures as a *valMatch()* function. Plots in figure 5.1 show the effect of the size of the contextual linkset $|LS(D_t, D_r)|$ on the overall count of evidence links $\sum I_1(G_{D_t}(e_t), G_{D_r}(e_r))$ and the number of distinct predicate-pair $|\bigcup I_2(G_{D_t}(e_t), G_{D_r}(e_r))|$. We fixed $\tau_{sem} = 0$ to see the effect of τ_{obj} on the evidence link count specifically. We then changed the value to $\tau_{sem} = 0.5$ to visualise the effect specifically on the distinct count of predicate-pairs that are considered as similar in the context of the application. For both experiments, we chose $\omega_{val} = 1$ to see the effect of each object similarity function as well. We notice that:

- in all of (a_1,a_2,a_3,a_4), the number of evidence links is proportional to the number of analyzed focus graphs. This points to the fact that focus graphs in both sides share a certain structure allowing to maintain a relatively fixed ratio of evidence links per pair of focus graphs. Moreover, in (a_1) compared to the absence of a threshold, more than half the evidence links were ignored in ($\tau_{obj} = 0.25$) indicating that those evidence links were of bad quality. As

[6] http://github.com/djebr/archer

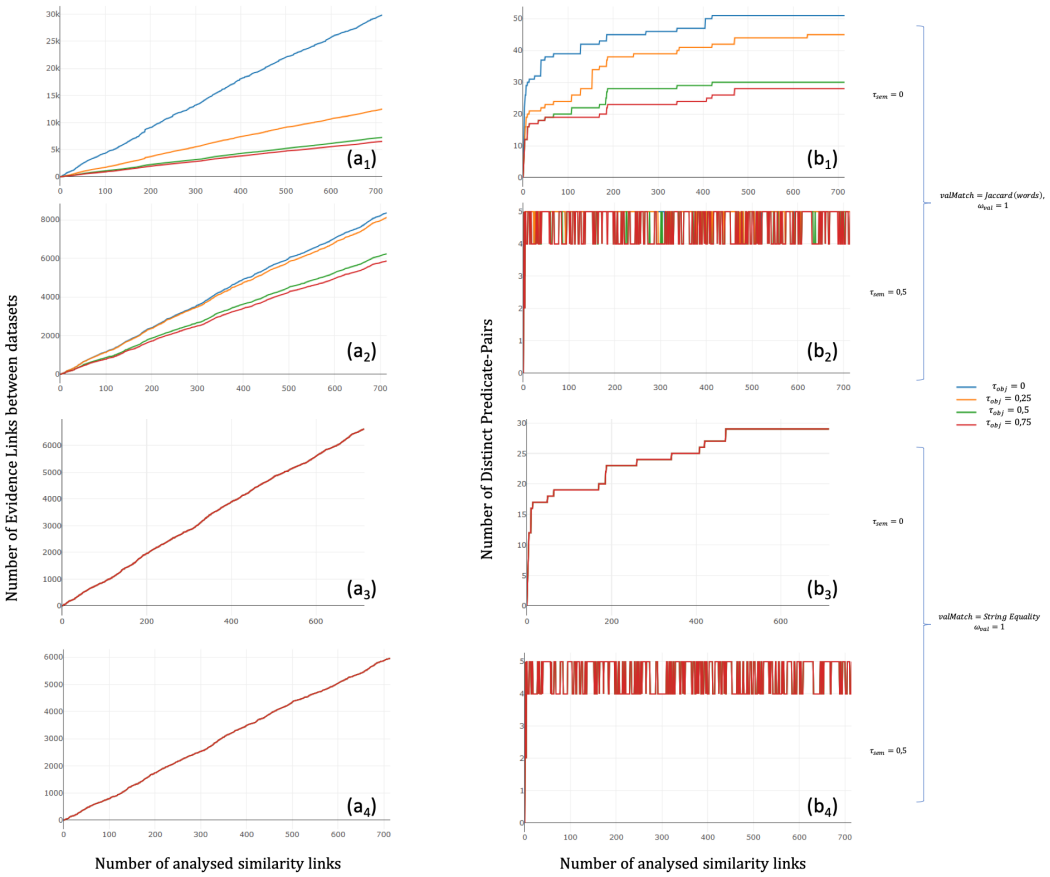

Fig. 5.1: Results for analysing 714 pairs of linked focus graphs: (a_i) total number of discovered evidence links (b_i) Number of distinct discovered predicate-pairs.

for the string equality, the local threshold is not needed as the indicators R_1 and R_3 will be the same (for each discovered evidence link, the quality is 1 at $\omega_{val} = 1$), so no evidence links will be dropped.

- as seen in (b_1,b_3), the effect of τ_{obj} on the number of evidence links is also predictable. The threshold will only allow links with better quality to be part of the overall evaluation.
- the number of predicate-pairs increases with the number of linked focus graphs. This is due to discovering predicate-pairs that did not have any evidence links in the first analyzed focus graphs. It is notable that for both object similarity methods, the number converges after analyzing more than 400 pairs of focus graphs. Furthermore, the effect of τ_{obj} can be observed

confirming that some predicate-pairs were dropped as they presented only bad quality links.
- when increasing τ_{sem}, the plots in (a_2) move closer to each other and converging towards (a_4) as it represents strict equality, resulting as well in similar shapes for (b_2) and (b_4). The fluctuation is due to the fact that the overall quality of some predicate-pairs evidence links might drops when considering new pairs of focus graphs that do not support the hypothesis. However, the plot remains constant proving that on the overall analysis, five predicate-pairs can be considered as best candidates to support the graph interlink.
- the difference between the number of predicate-pairs in (b_3) and (b_4) is remarkable. Comparing to 28 predicate-pairs in the first with 6600 evidence links, the second has only 5 predicate-pairs with almost 6000 evidence links. This further provides proof that most of the discovered links were not of general use (not common information between focus graph pairs).

6 Discussion

An argument about statistical extraction of semantics would be the fact that a target dataset can be completely wrong, or somehow unrelated to the reference dataset like having the same information but in a different language. In both cases, this does not affect the semantic analysis of evidence links. For the first case, no links will be discovered and this will raise a flag about the current configuration itself (one is wrong about everything related to a certain subject, or the references were chosen incorrectly). For the second case, the similarity links will not be translated as well, and triggering the intuition of completeness between the two graphs (and not that of negation).

Analyzing the similarity patterns based on graph interlinks may be a good first base to evaluate trustworthiness and inclusion between data sources. This approach works best if one has already a clustered dataset by structure, and the system is used to see the reliability of its information according to known sources.

Further investigations are scheduled to explore the use of other clustering methods, or customized focus graphs and see the possibility to transform the existing information about focus graphs using graph embedding. Finally, user queries should be one of the main triggers of uncertainty measurement, interlink creation, and evaluation.

7 Conclusion

We have proposed an approach to evaluate uncertainty in linked data sources by providing graph interlinks. Our approach is based on both object and predicate similarity and operates on different levels to evaluate task-specific uncertainty measurements for the data source of interest. The results of our experiments show that graph interlinks can be supported with a set of evidence links, depending on the use-case and the user's choice of quality parameters. Using our tool enables us

to assess the quality of a dataset regarding a certain task, and annotate its data accordingly while producing reusable and publishable uncertainty measurements. Future work will focus on learning the most suitable structure for a focus graph, generalizing our approach to consider a set of reference resources, and including more parameters such as scores from ontology matchers.

References

1. Gandon, F.: Web Science, Artificial Intelligence and Intelligence Augmentation (in Dagstuhl Perspectives Workshop 18262 - 10 Years of Web Science: Closing The Loop). Other, Dagstuhl (2019), https://hal.inria.fr/hal-01976768
2. Djebri, A.E.A., Tettamanzi, A.G.B., Gandon, F.: Publishing Uncertainty on the Semantic Web: Blurring the LOD Bubbles. In: Endres, D., Alam, M., Şotropa, D. (eds.) Graph-Based Representation and Reasoning. pp. 42–56. Springer International Publishing, Cham (2019)
3. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web **8**(3), 489–508 (2017)
4. Paulheim, H.: Identifying wrong links between datasets by multi-dimensional outlier detection. In: WoDOOM. pp. 27–38 (2014)
5. Hogan, A., Polleres, A., Umbrich, J., Zimmermann, A.: Some entities are more equal than others: statistical methods to consolidate linked data. In: 4th International Workshop on New Forms of Reasoning for the Semantic Web: Scalable and Dynamic (NeFoRS2010) (2010)
6. Rico, M., Mihindukulasooriya, N., Kontokostas, D., Paulheim, H., Hellmann, S., Gómez-Pérez, A.: Predicting incorrect mappings: a data-driven approach applied to DBpedia. In: Proceedings of the 33rd annual ACM symposium on applied computing. pp. 323–330 (2018)
7. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. In: Journal on data semantics IV, pp. 146–171. Springer (2005)
8. Fernandes, A.A., Paton, N.W.: Quantifying and Propagating Uncertainty in Automated Linked Data Integration. Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVII **10940**, 81 (2018)
9. Rico, M., Mihindukulasooriya, N., Kontokostas, D., Paulheim, H., Hellmann, S., Gómez-Pérez, A.: Predicting incorrect mappings: a data-driven approach applied to DBpedia. In: Proceedings of the 33rd annual ACM symposium on applied computing. pp. 323–330 (2018)
10. Cheatham, M., Cruz, I.F., Euzenat, J., Pesquita, C.: Special issue on ontology and linked data matching. Semantic Web **8**(2), 183–184 (2017). https://doi.org/10.3233/SW-160251, https://doi.org/10.3233/SW-160251
11. Brickley, D., Guha, R.V., McBride, B.: RDF Schema 1.1. W3C recommendation **25**, 2004–2014 (2014)
12. McGuinness, D.L., Van Harmelen, F., et al.: OWL web ontology language overview. W3C recommendation **10**(10), 2004 (2004)
13. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation **18**, W3C (2009)